LDA SE Tutorial

Abram Hindle

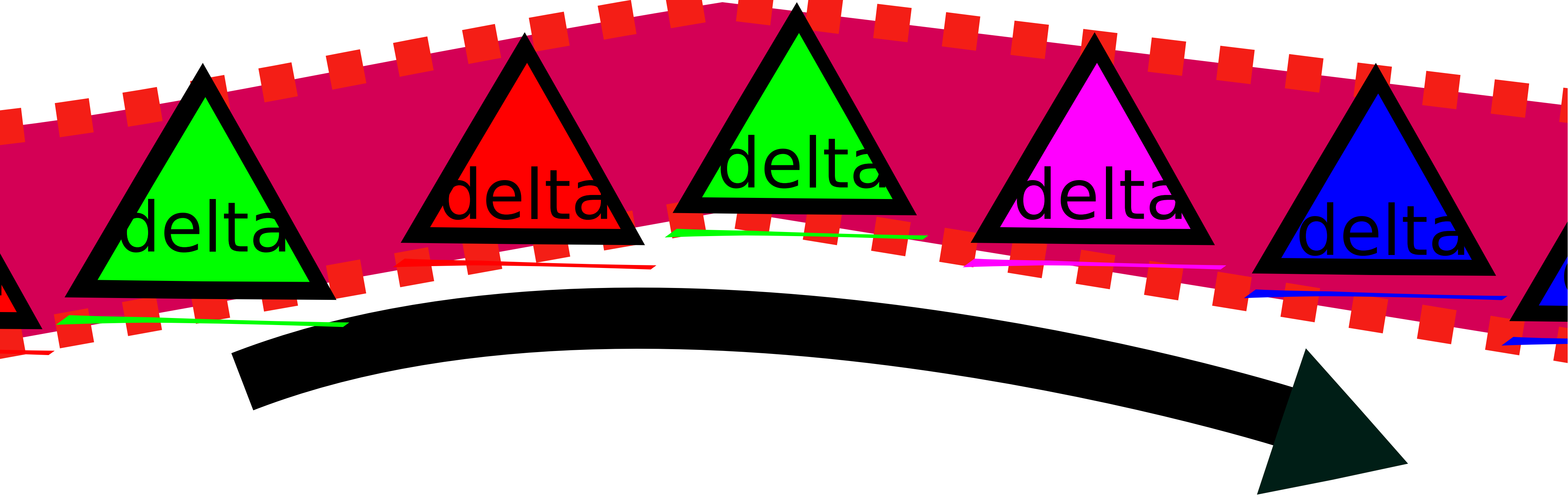
<abram.hindle@ualberta.ca>

Department of Computing Science

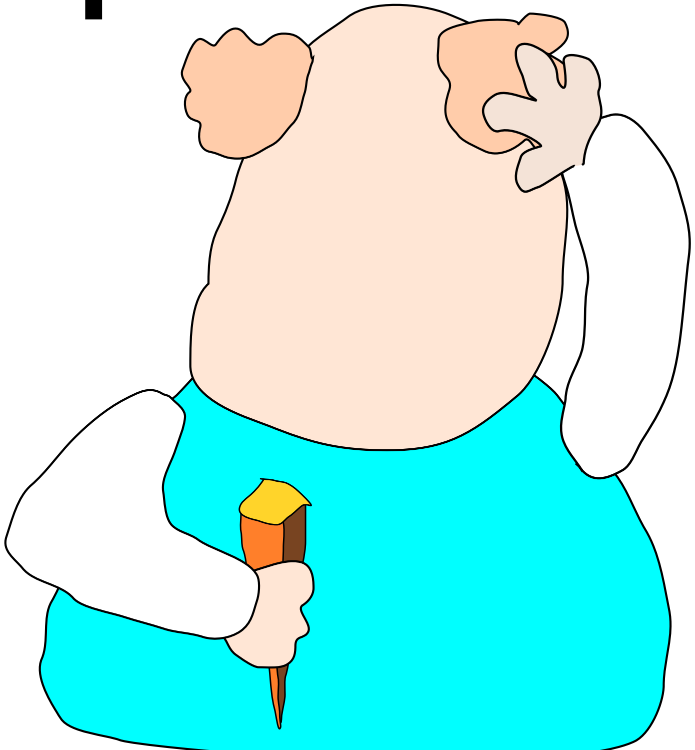
University of Alberta

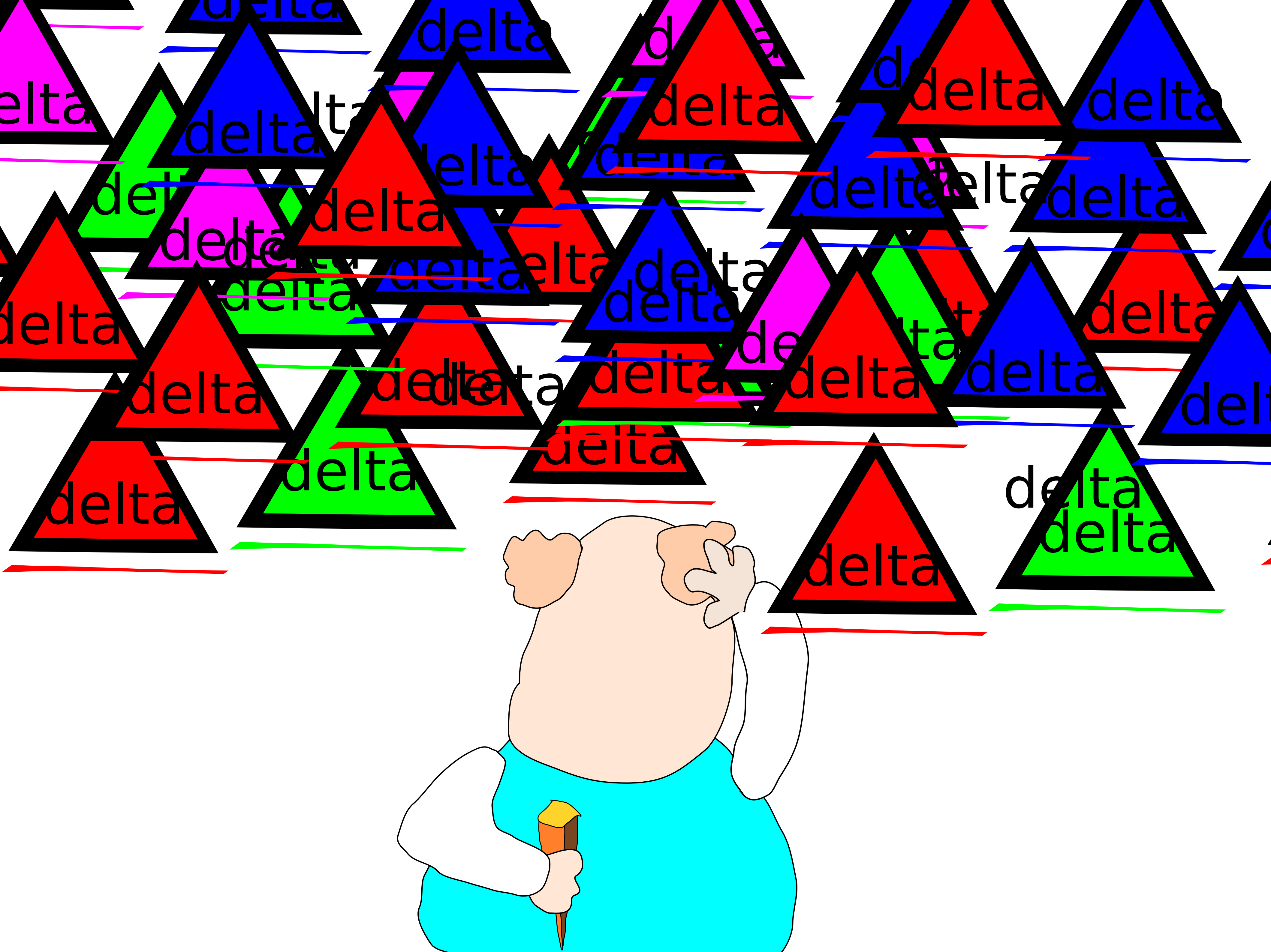
Edmonton, AB, Canada

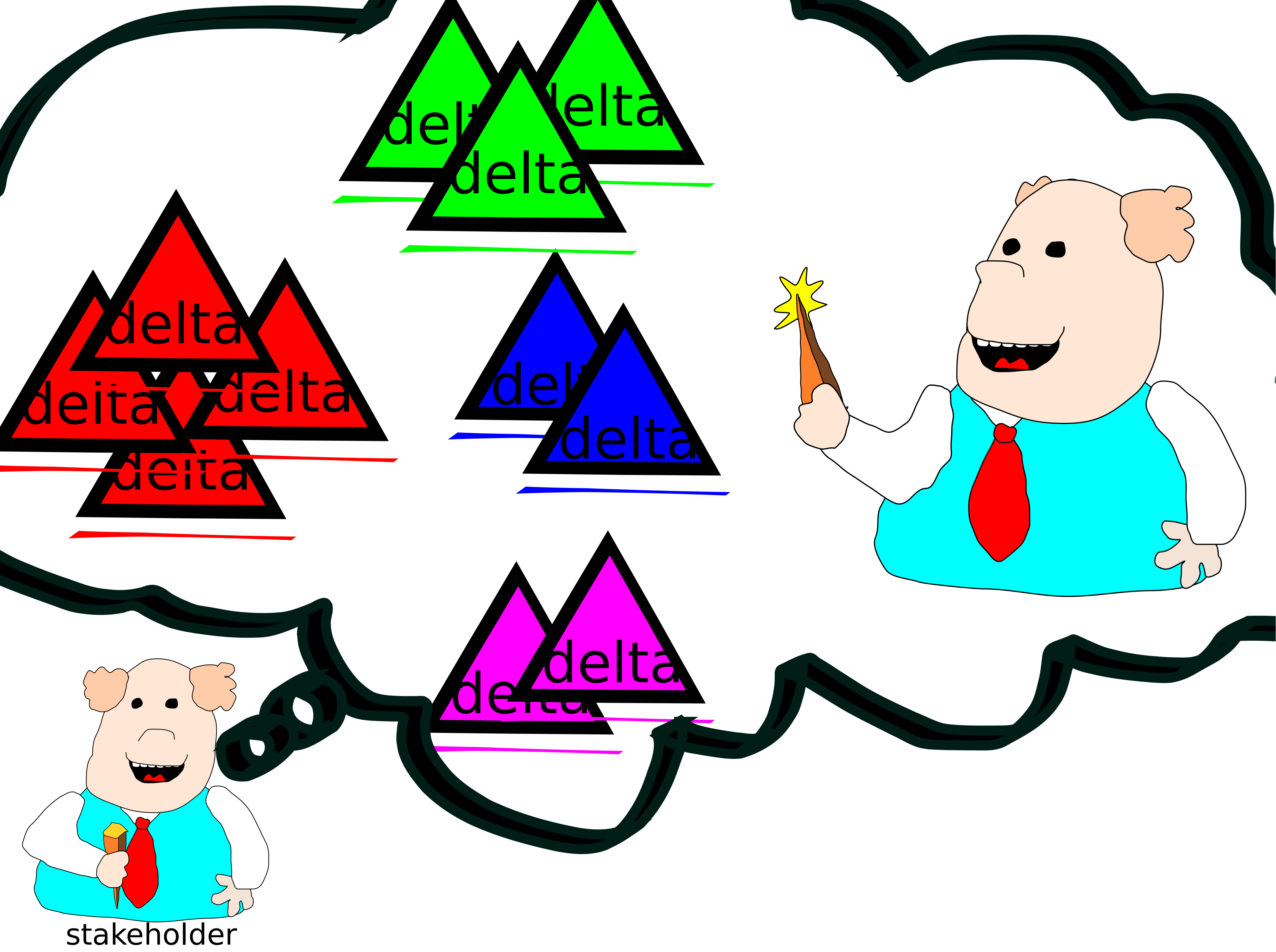
http://softwareprocess.ca



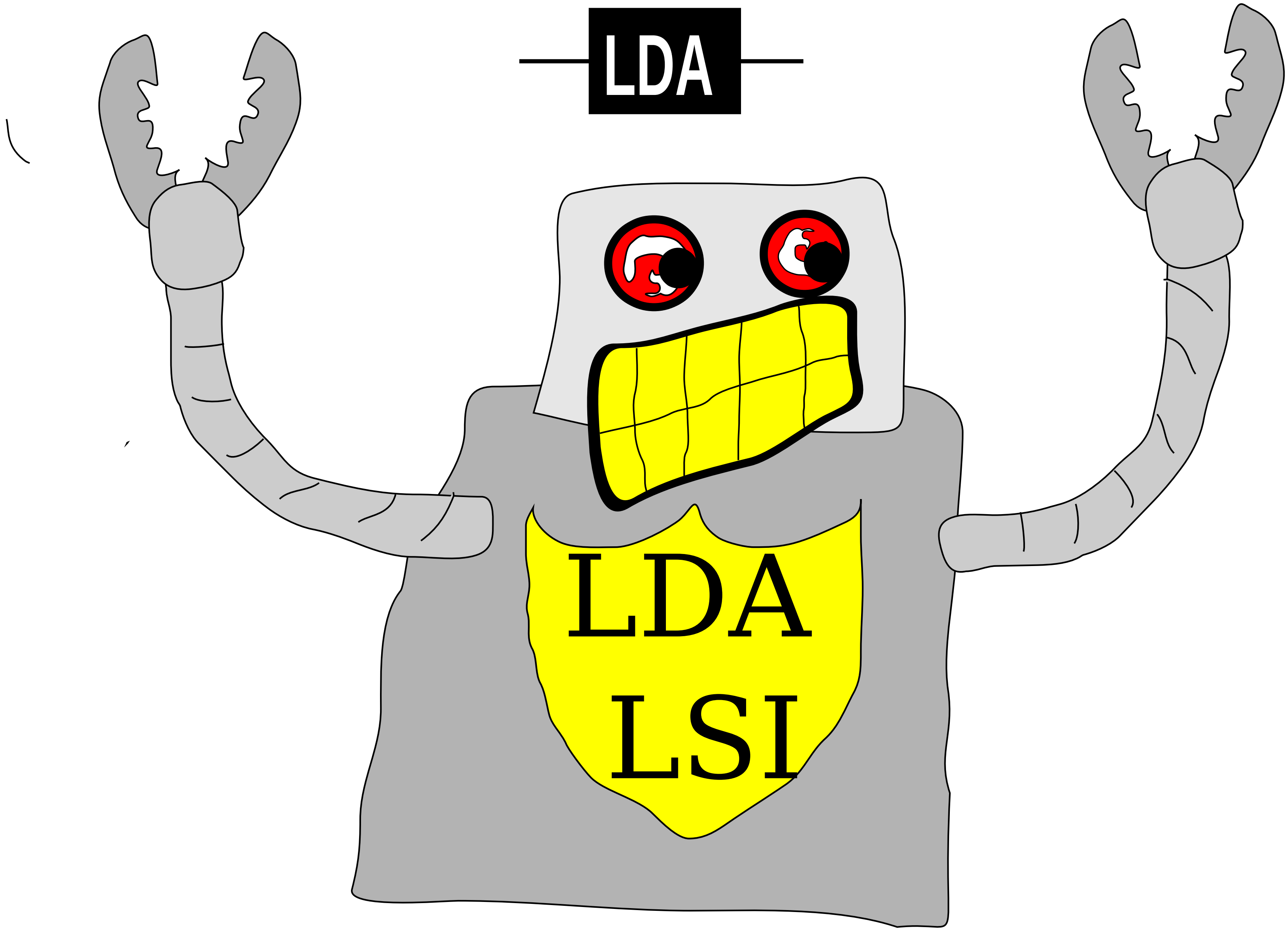
Development History



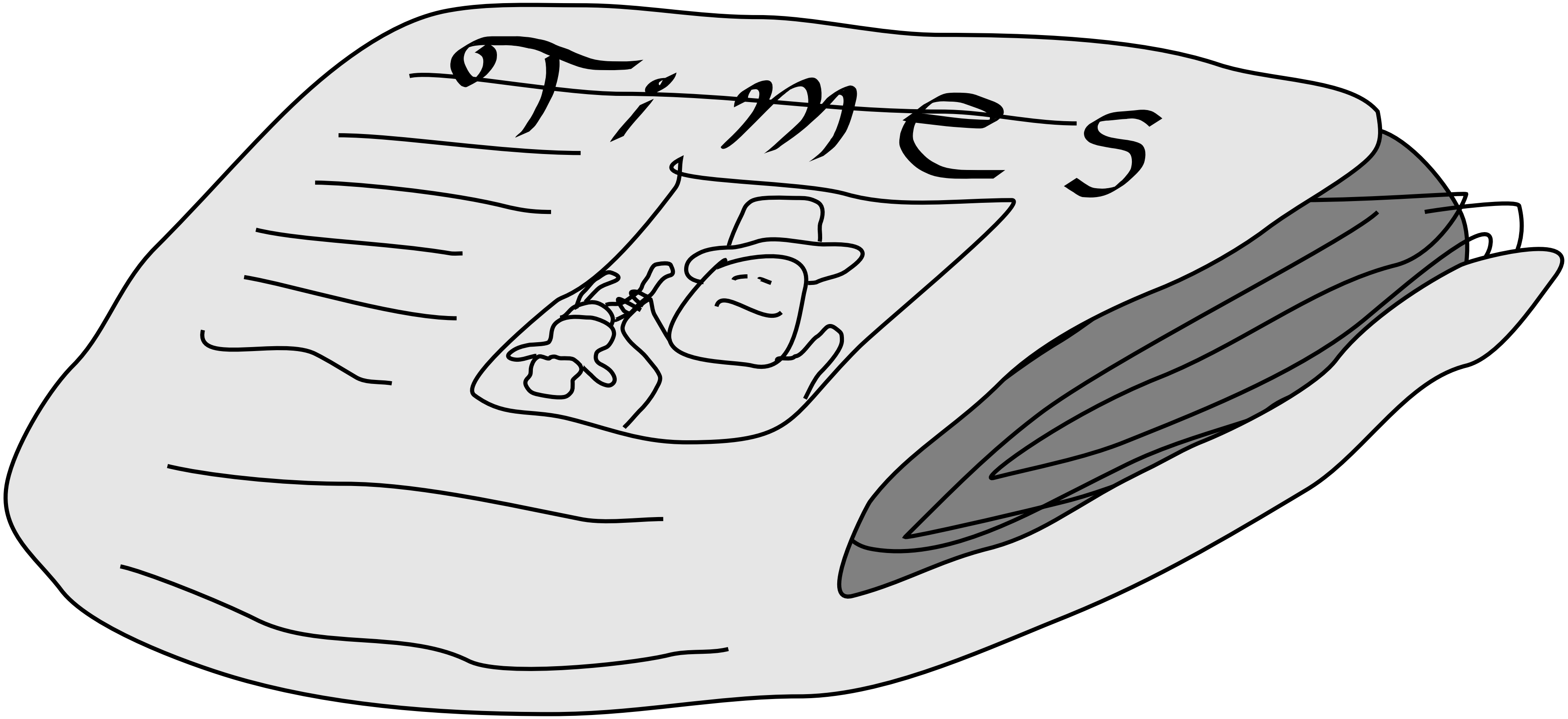

delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta
delta


delt
delta
delta
delta
deita
delta
deita
del
delta
delta
de
stakeholder

Our blackbox

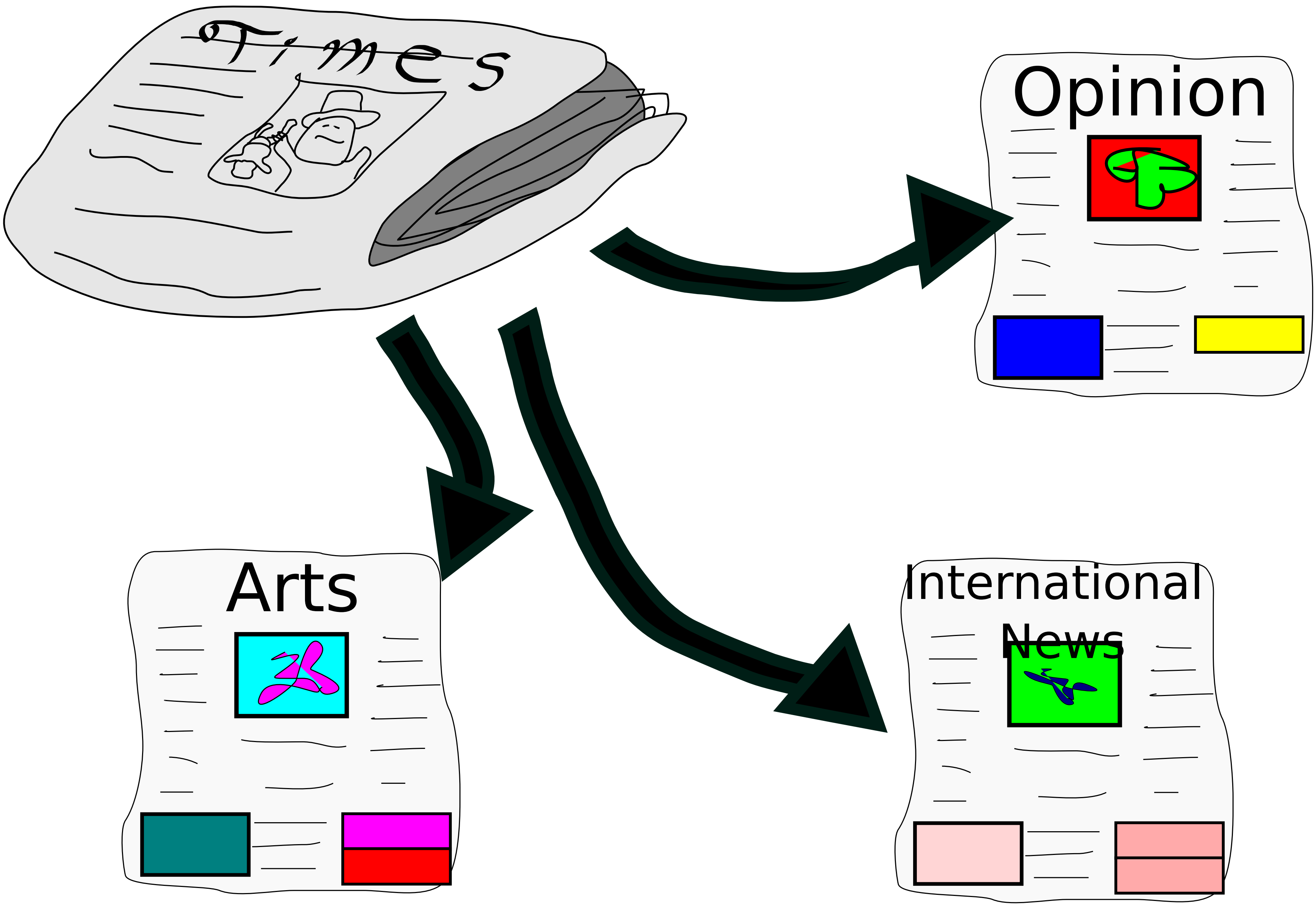


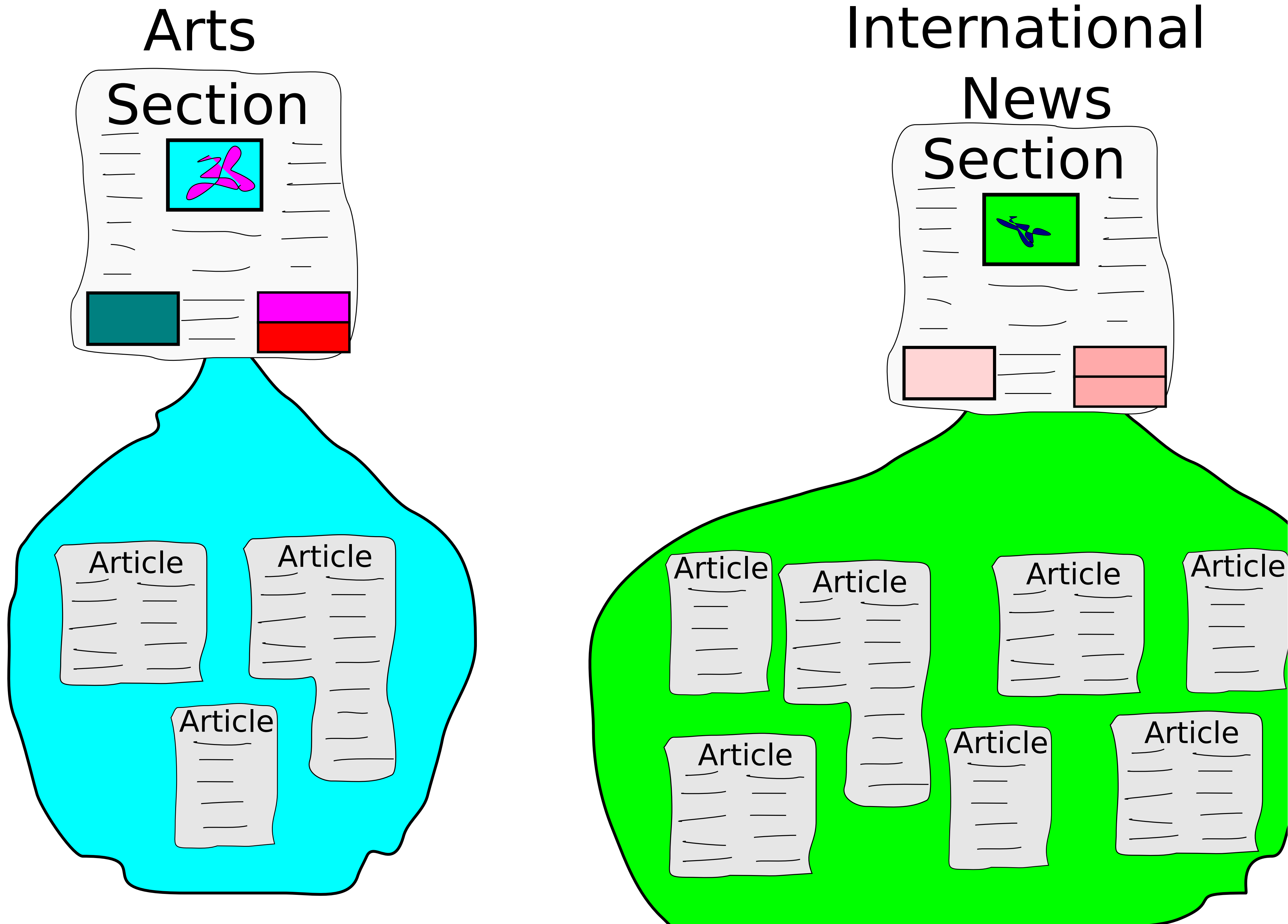
Example



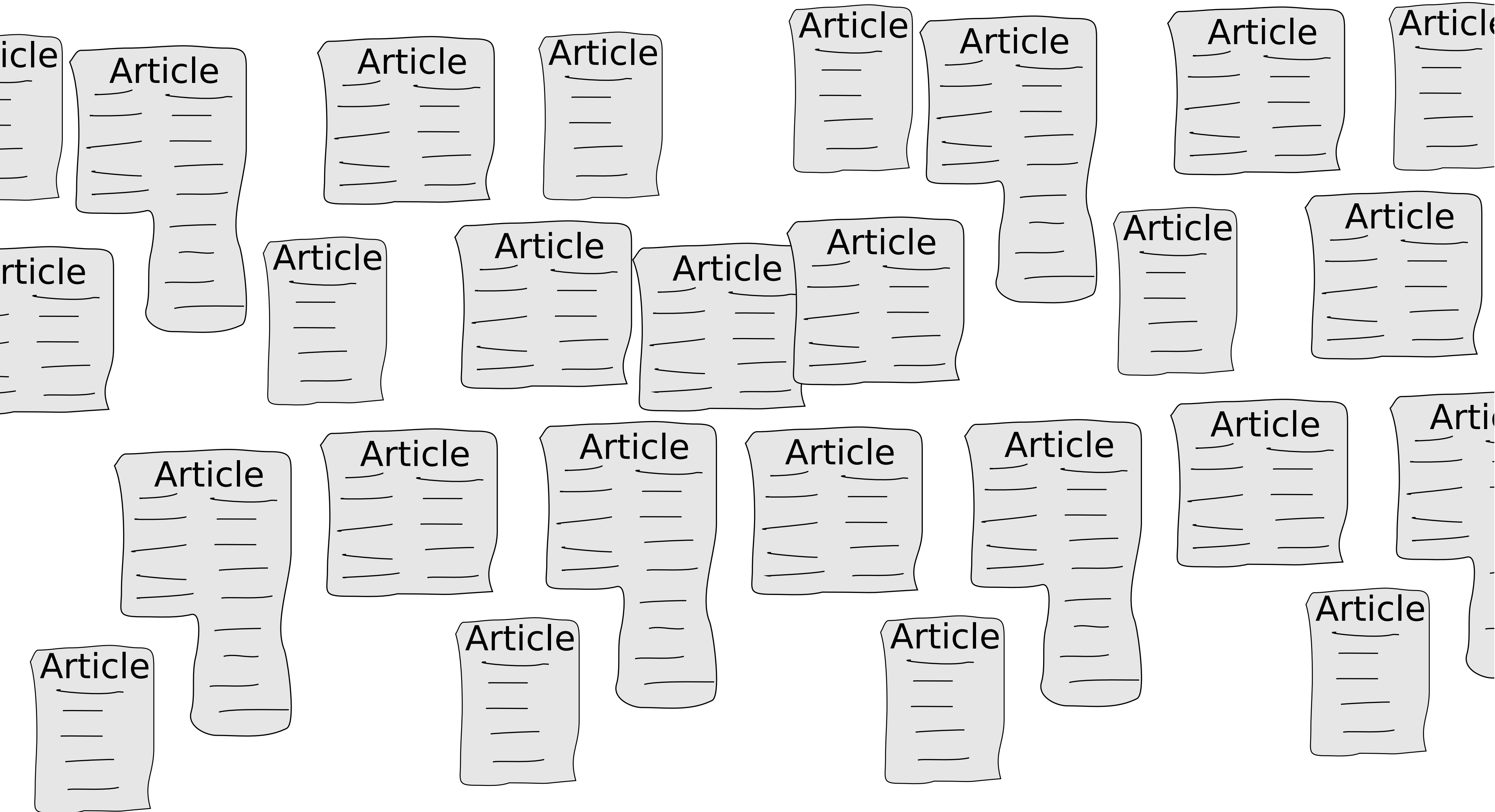
[Blei]

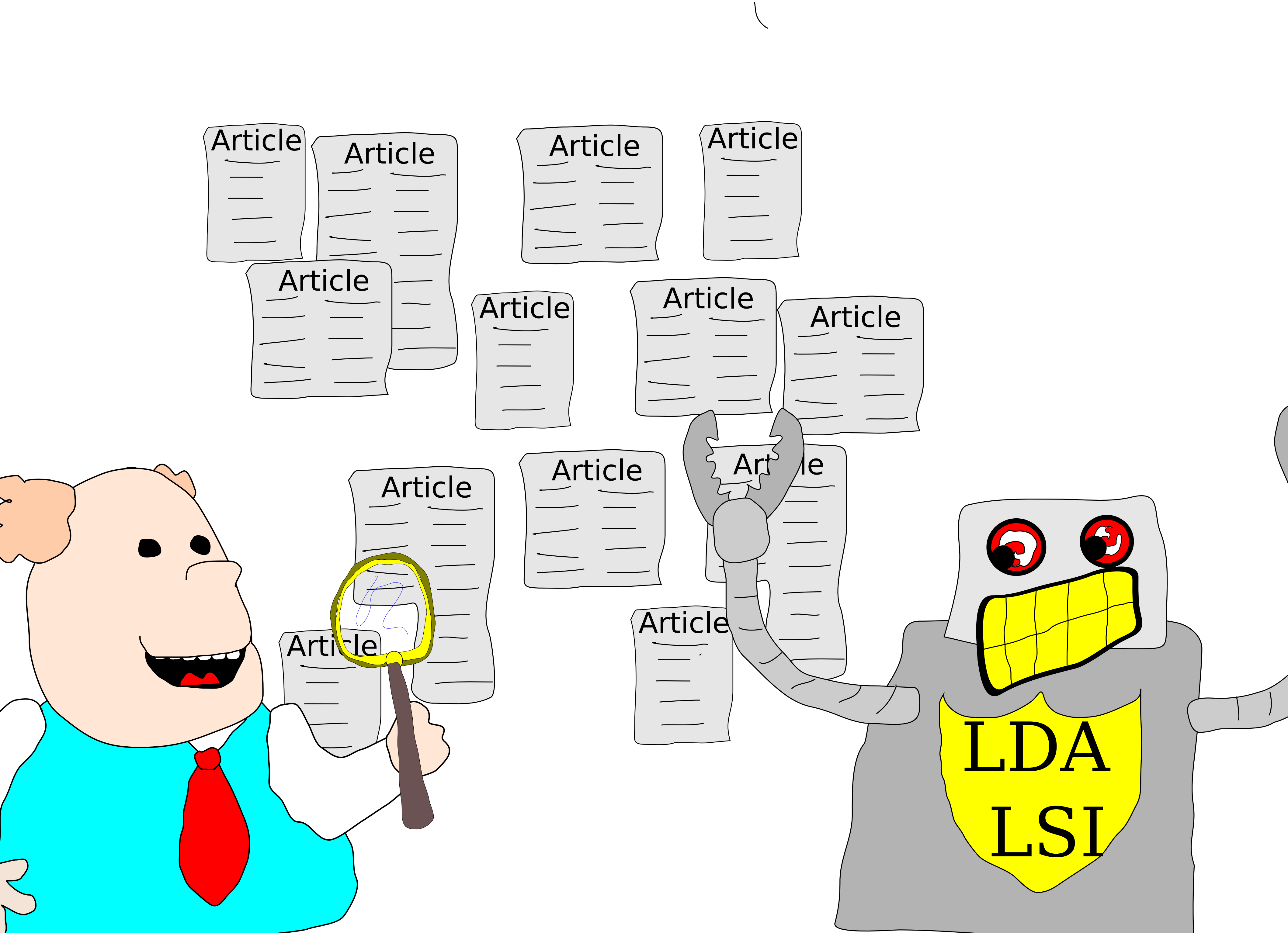
apologies to those with prior LDA/LSI experience

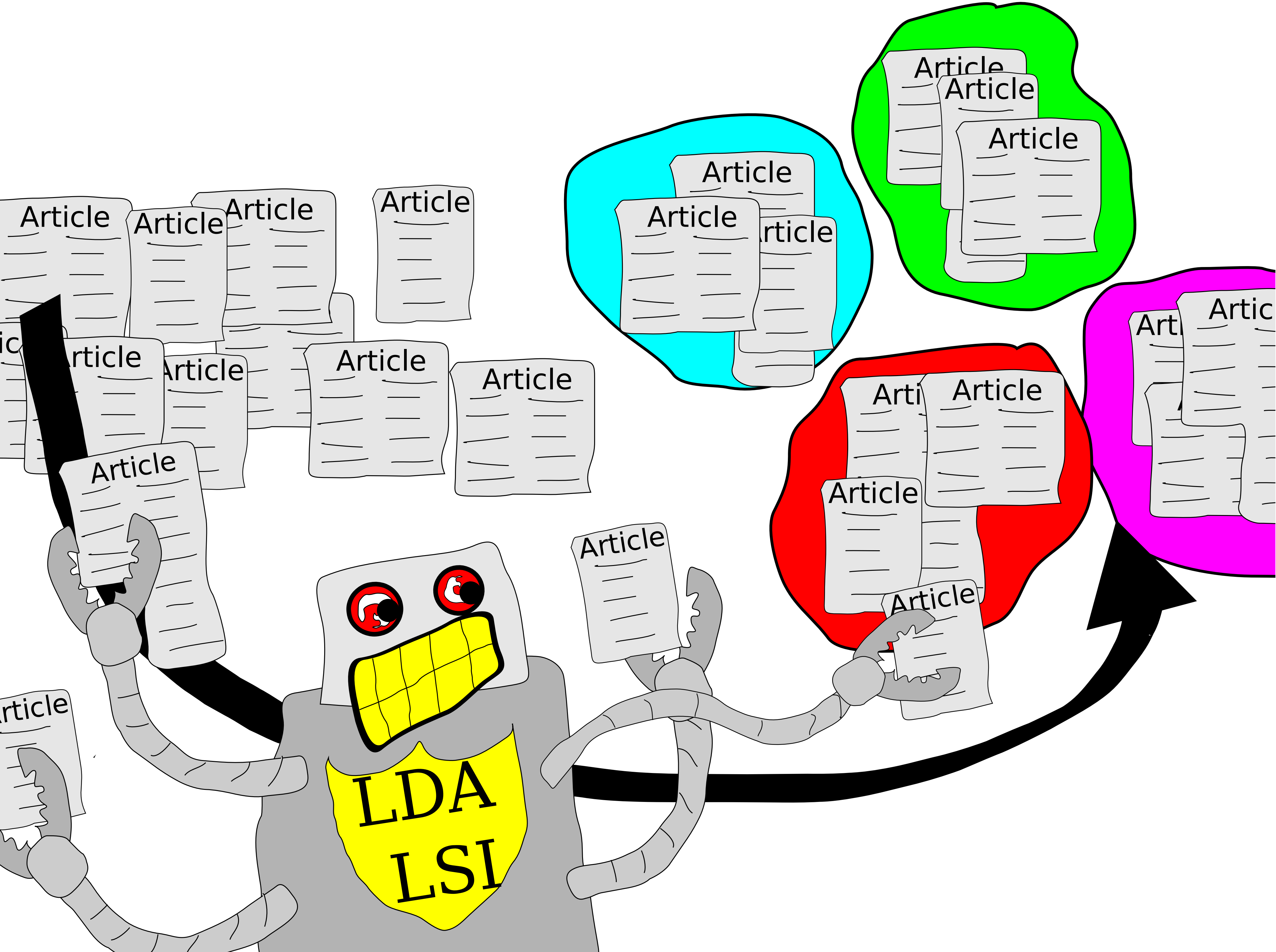

Times
Opinion
Arts
International News

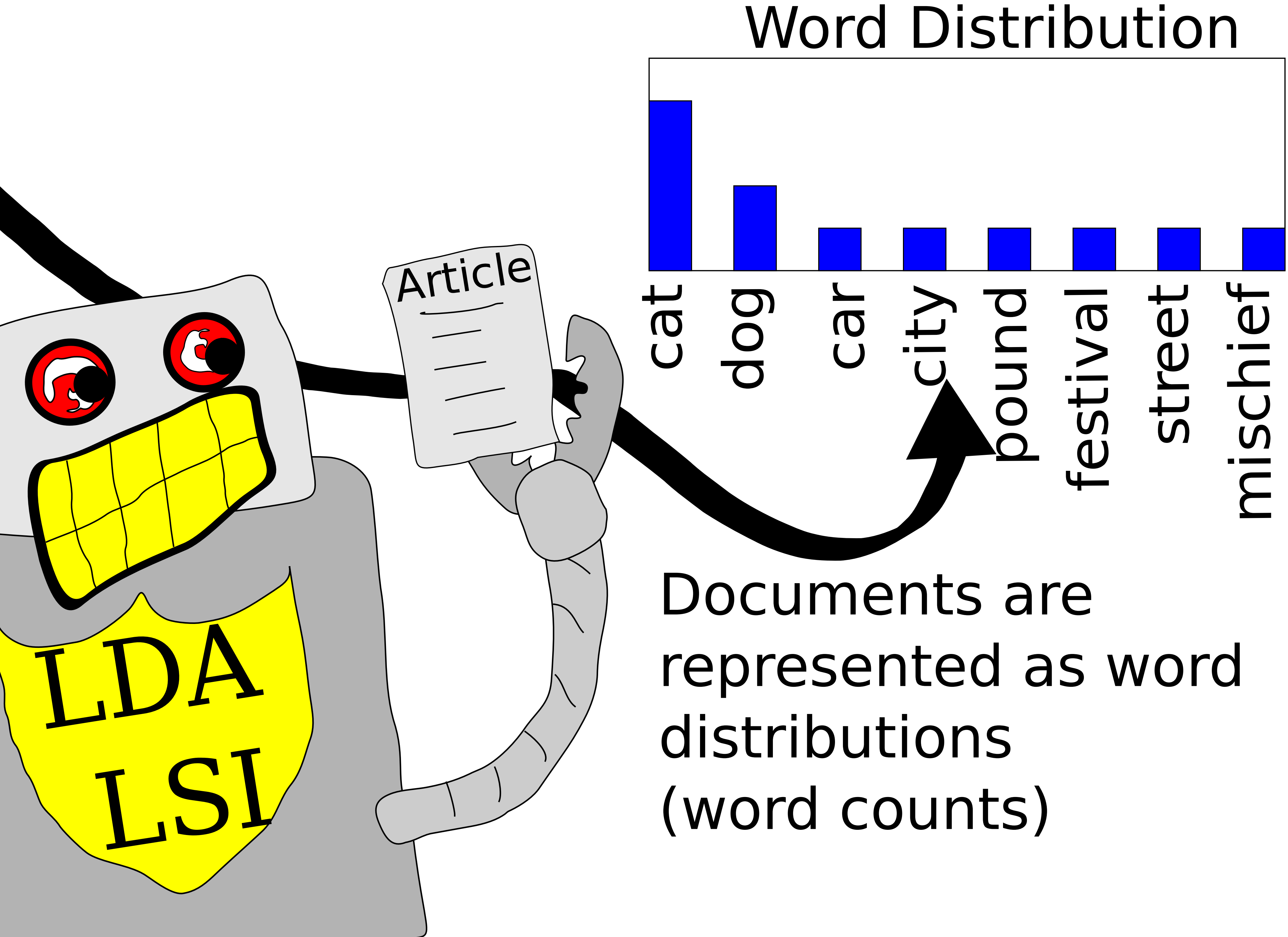

Arts
Section
International
News
Section
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article

What if we didn't know what section the articles were in?



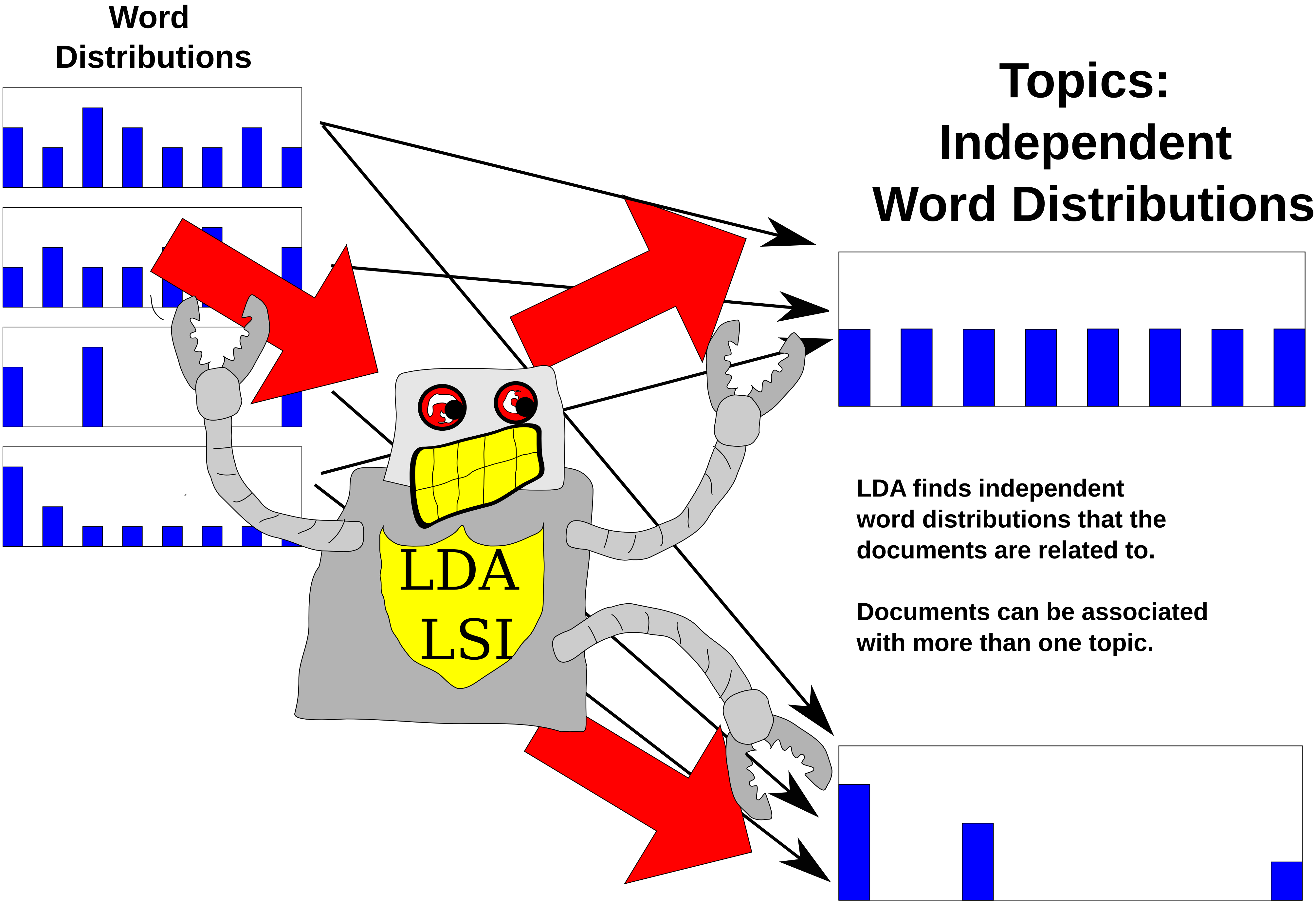

Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
LDA
LSI


Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
Article
rticle
Article
Article
Article
Article
Article
Article
Arti
Article
Article
Article
Arti
Artic
rticle
LDA
LSI



Documents are represented as word distributions (word counts)

Topics: Independent Word Distributions



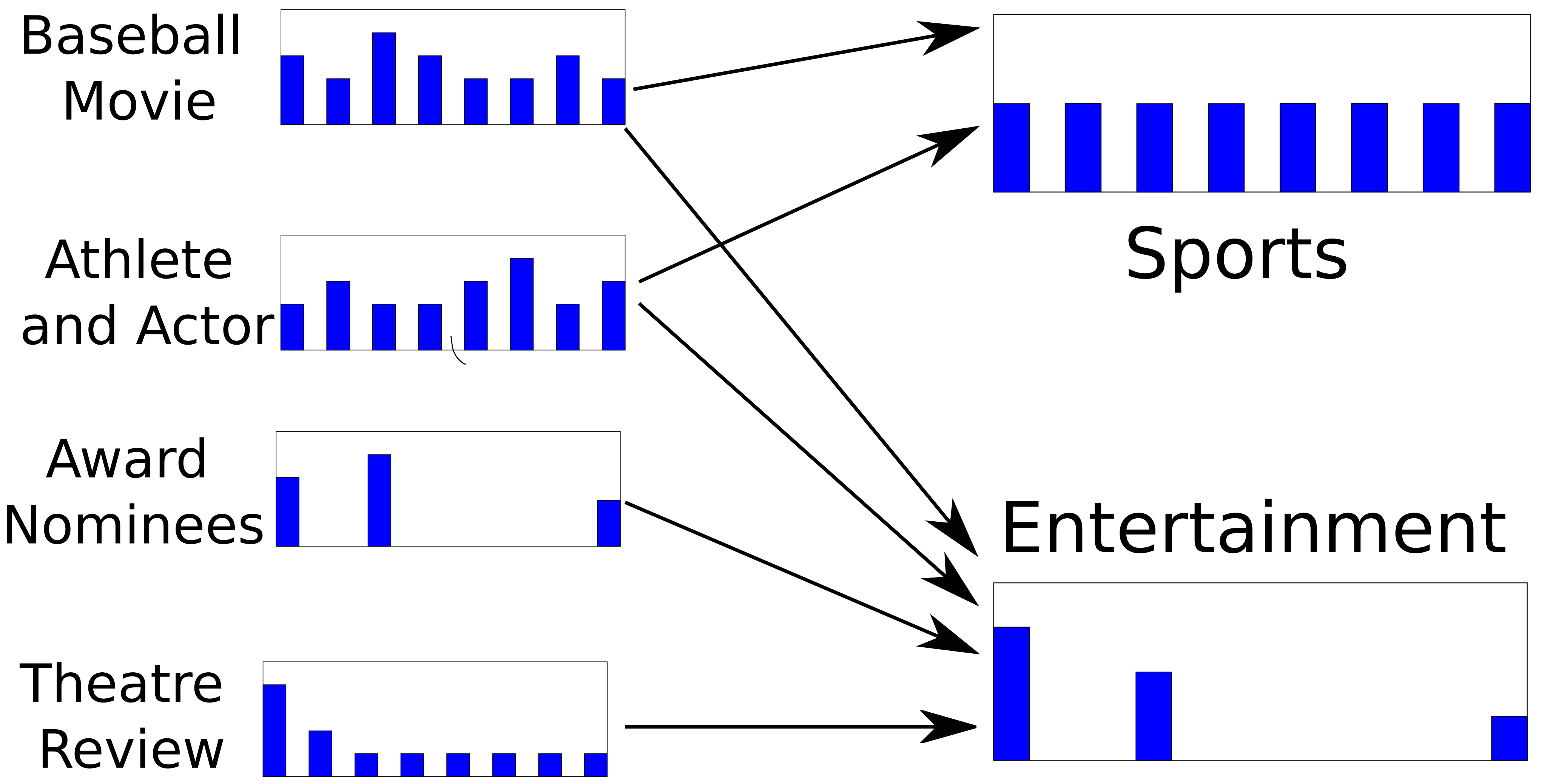
LDA finds independent word distributions that the documents are related to.

Documents can be associated with more than one topic.

Original Article

Word Distributions

Topics: Independent Word Distributions



Baseball Movie

Athlete and Actor

Award Nominees

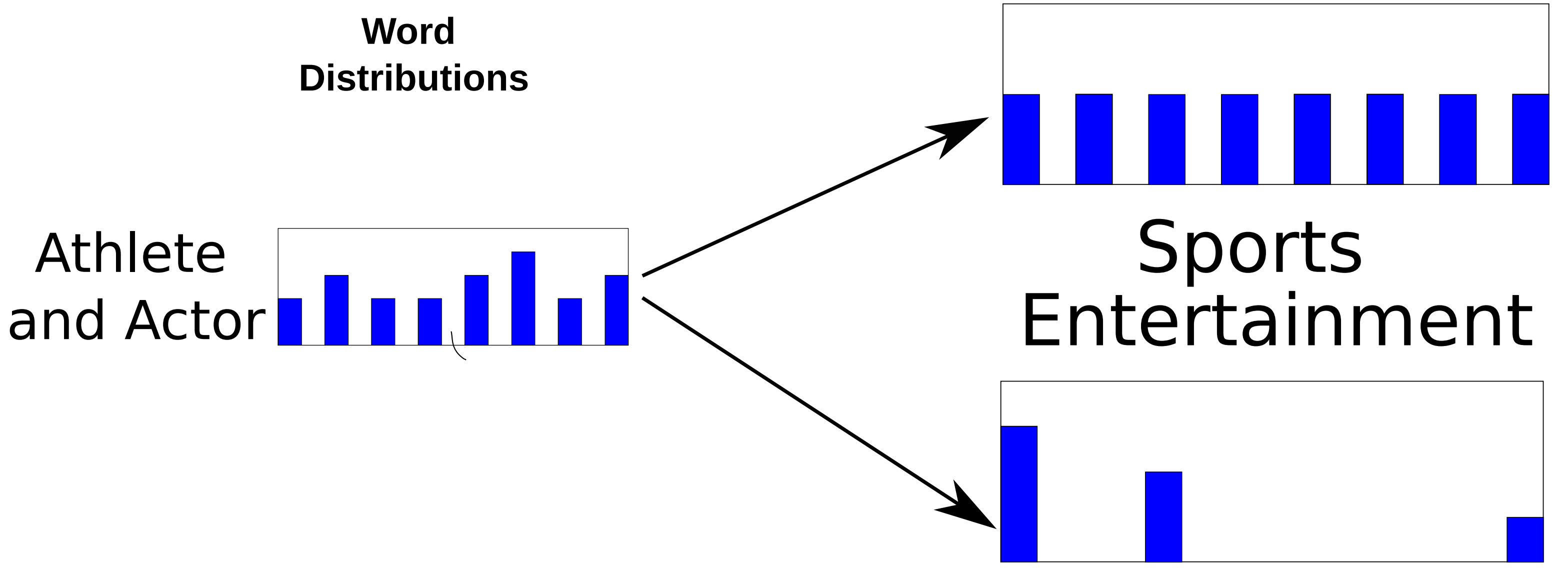
Theatre Review

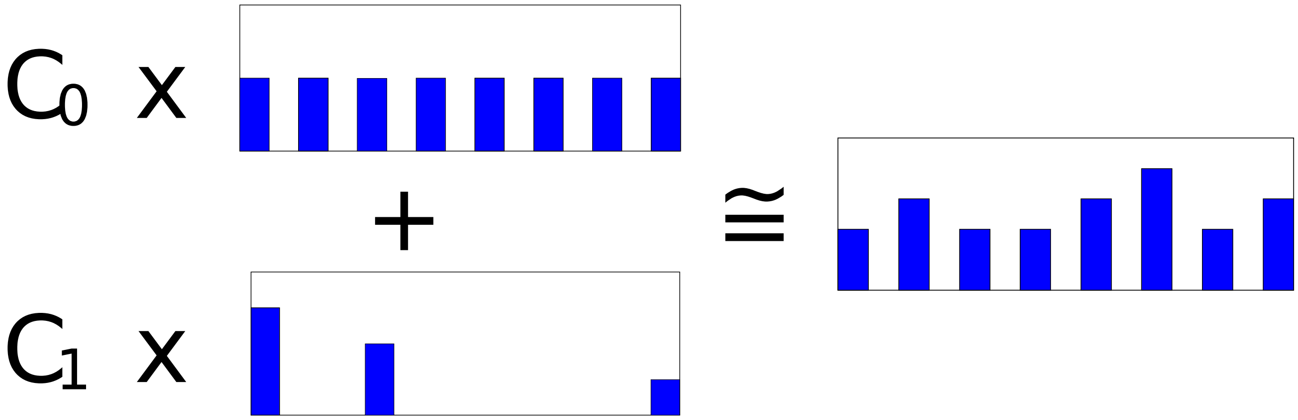
Sports

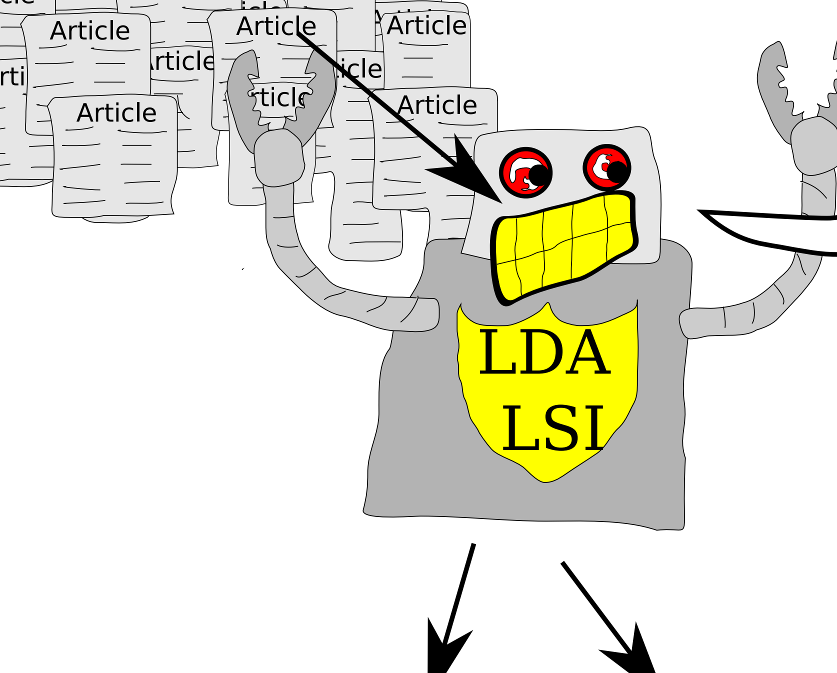
Entertainment

Documents are allocated to topics and proportion of their words are allocated to a topic. Because it is allocation, it means that topics share limited words or allocations. You can't have two topics allocated at 100% to one document.

Topics:
Independent
Word Distributions







Here are two topics. I don't know what they are about!

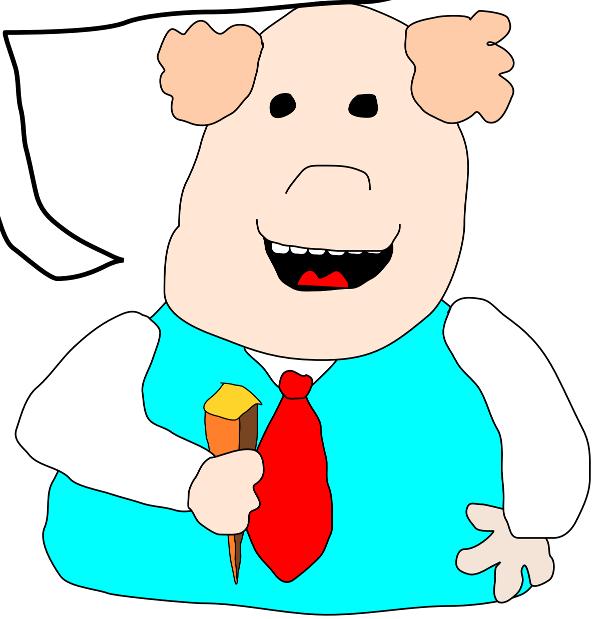
Topic 1

* play
* game
* inning
* player
* quarter
* opponent
* ...

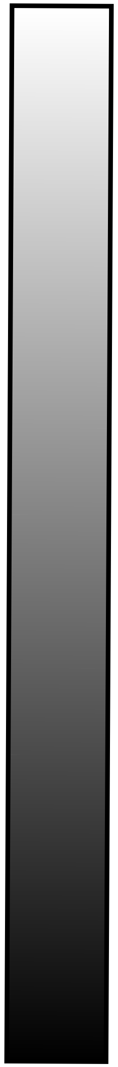
Topic 2

* gambling
* play
* night life
* comedy
* movie
* theatre
* ...

These word lists look look like: **Sports** an **Entertainment** !



Many Documents



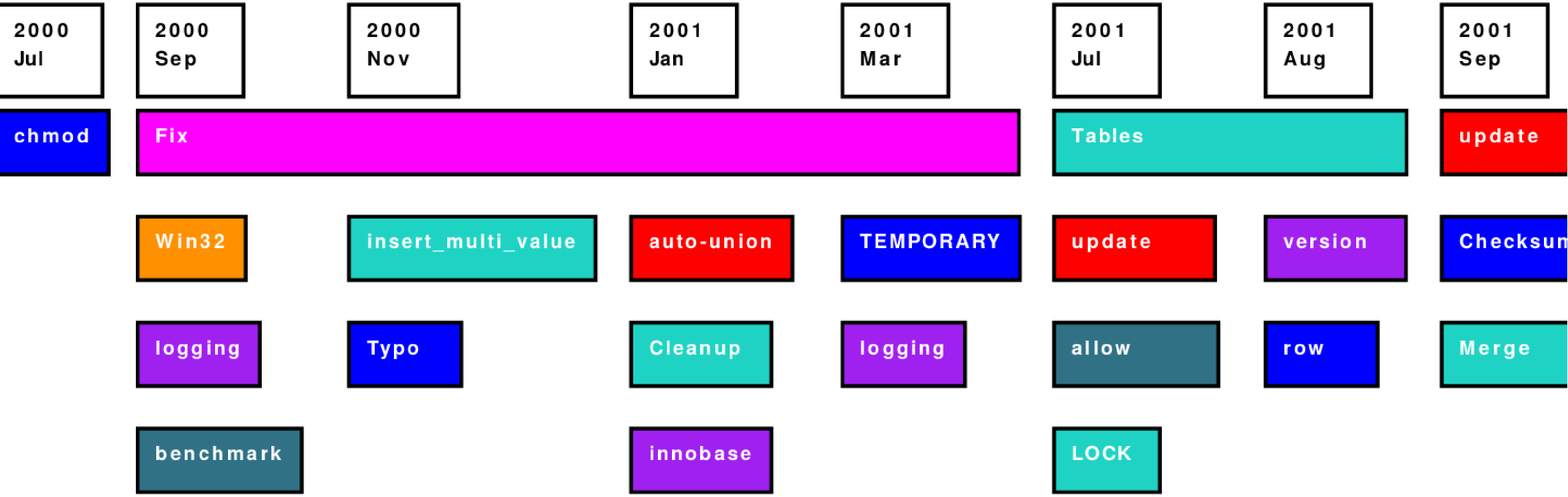
Few Documents

Topic 1

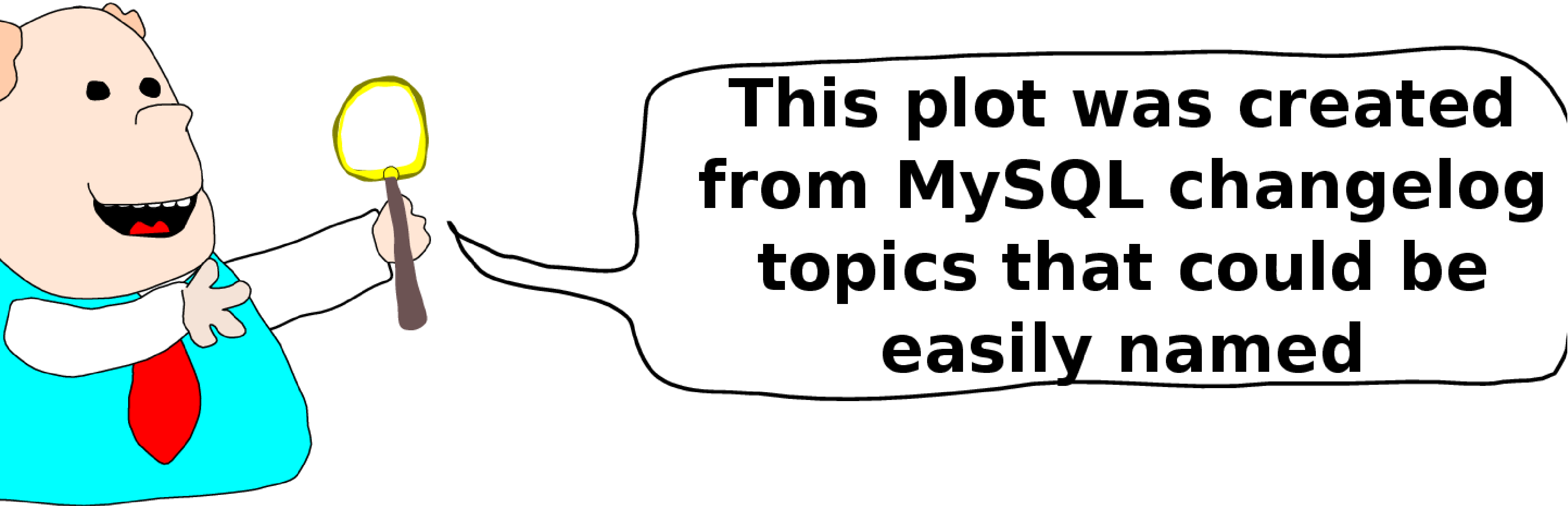
Topic 10

Topic 20

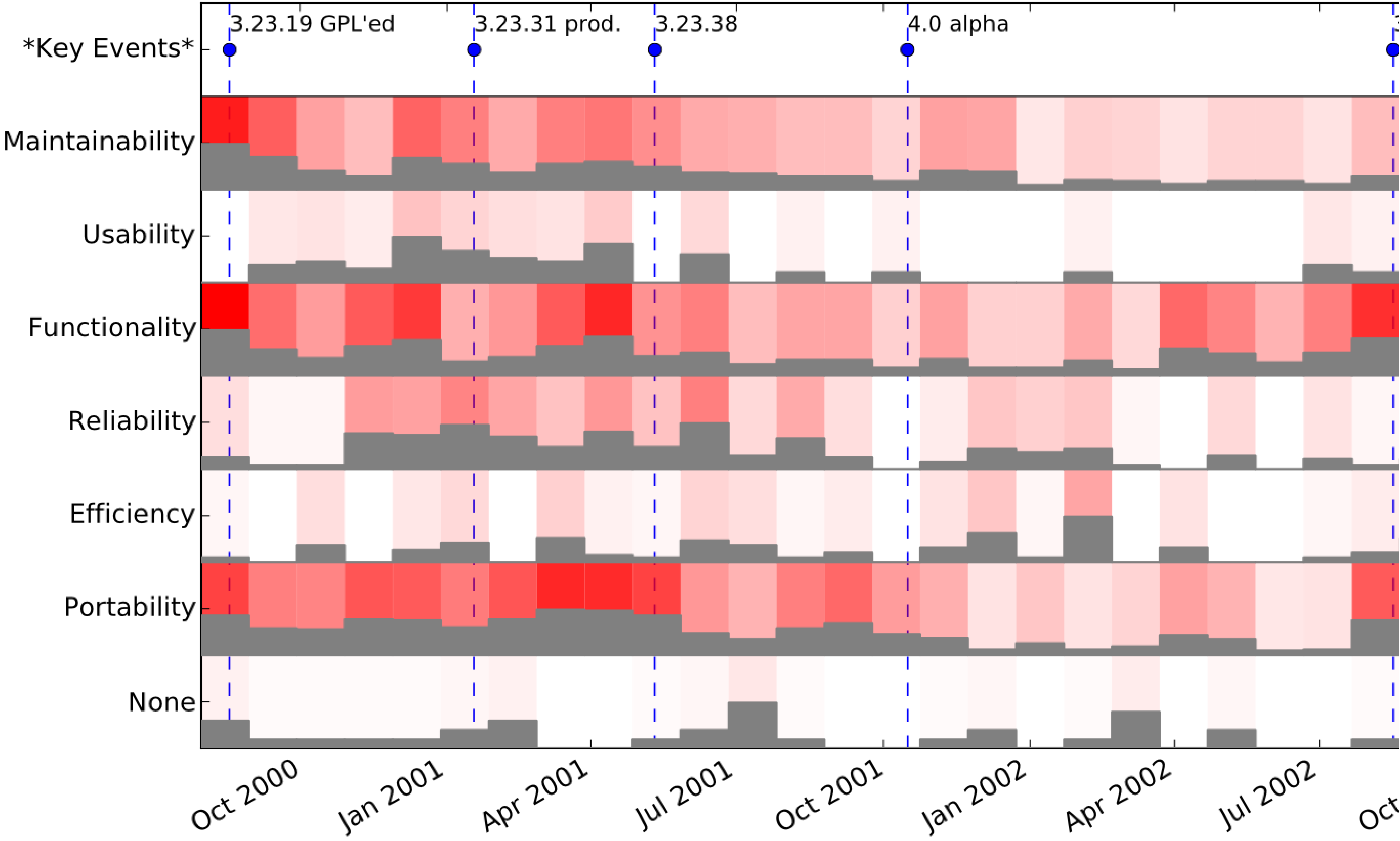
MySQL 3.23 Case Study



2000 Jul	2000 Sep	2000 Nov	2001 Jan	2001 Mar	2001 Jul	2001 Aug	2001 Sep
chmod	Fix				Tables		update
	Win32	insert_multi_value	auto-union	TEMPORARY	update	version	Checksum
	logging	Typo	Cleanup	logging	allow	row	Merge
	benchmark		innobase		LOCK		


This plot was created from MySQL changelog topics that could be easily named

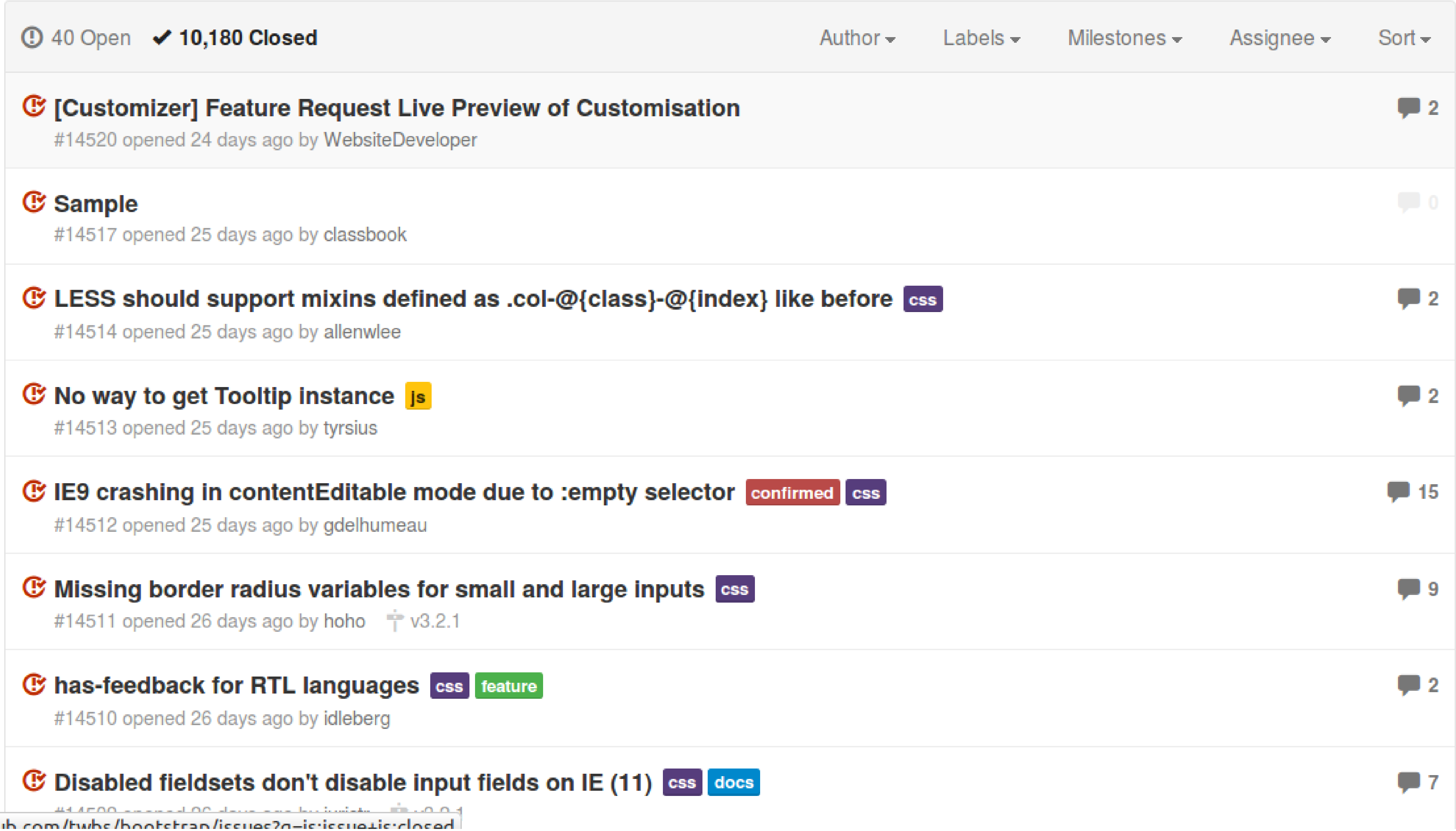
2004 Jun
2004 Jul
2004 Aug
2004 Sep
2004 Oct
2004 Nov
2004 Dec
2005 Jan
2005 Feb
2005 Mar
2005 Apr
2005 May
2005 Jun
2005 Jul
2005 Aug
2005 Sep
2005 Oct
2005 Nov
2005 Dec
2006 Jan
2006 Feb
2006 Mar
2006 Apr
2006 May
2006 Jun


Key Events
3.23.19 GPL'ed
3.23.31 prod.
3.23.38
4.0 alpha
Maintainability
Usability
Functionality
Reliability
Efficiency
Portability
None
Oct 2000
Jan 2001
Apr 2001
Jul 2001
Oct 2001
Jan 2002
Apr 2002
Jul 2002
Oct

Data

- Choose:
  - Source Code
  - Natural Language
    - You can mix the two but you're going to bias topics to either language.
    - Try to stick to 1 natural language. If you have a primarily English project the German contributors will be noticable.
- Need to tokenize/split words
- Blei does not recommend n-grams but you don't need to listen to him. He just made LDA that's all.

Data: Issue Trackers



ⓘ 40 Open ✓ **10,180 Closed**

[Customizer] Feature Request Live Preview of Customisation 💬 2
#14520 opened 24 days ago by WebsiteDeveloper

Sample 💬 0
#14517 opened 25 days ago by classbook

LESS should support mixins defined as .col-@{class}-@{index} like before css 💬 2
#14514 opened 25 days ago by allenwlee

No way to get Tooltip instance js 💬 2
#14513 opened 25 days ago by tyrsius

IE9 crashing in contentEditable mode due to :empty selector confirmed css 💬 15
#14512 opened 25 days ago by gdelhumeau

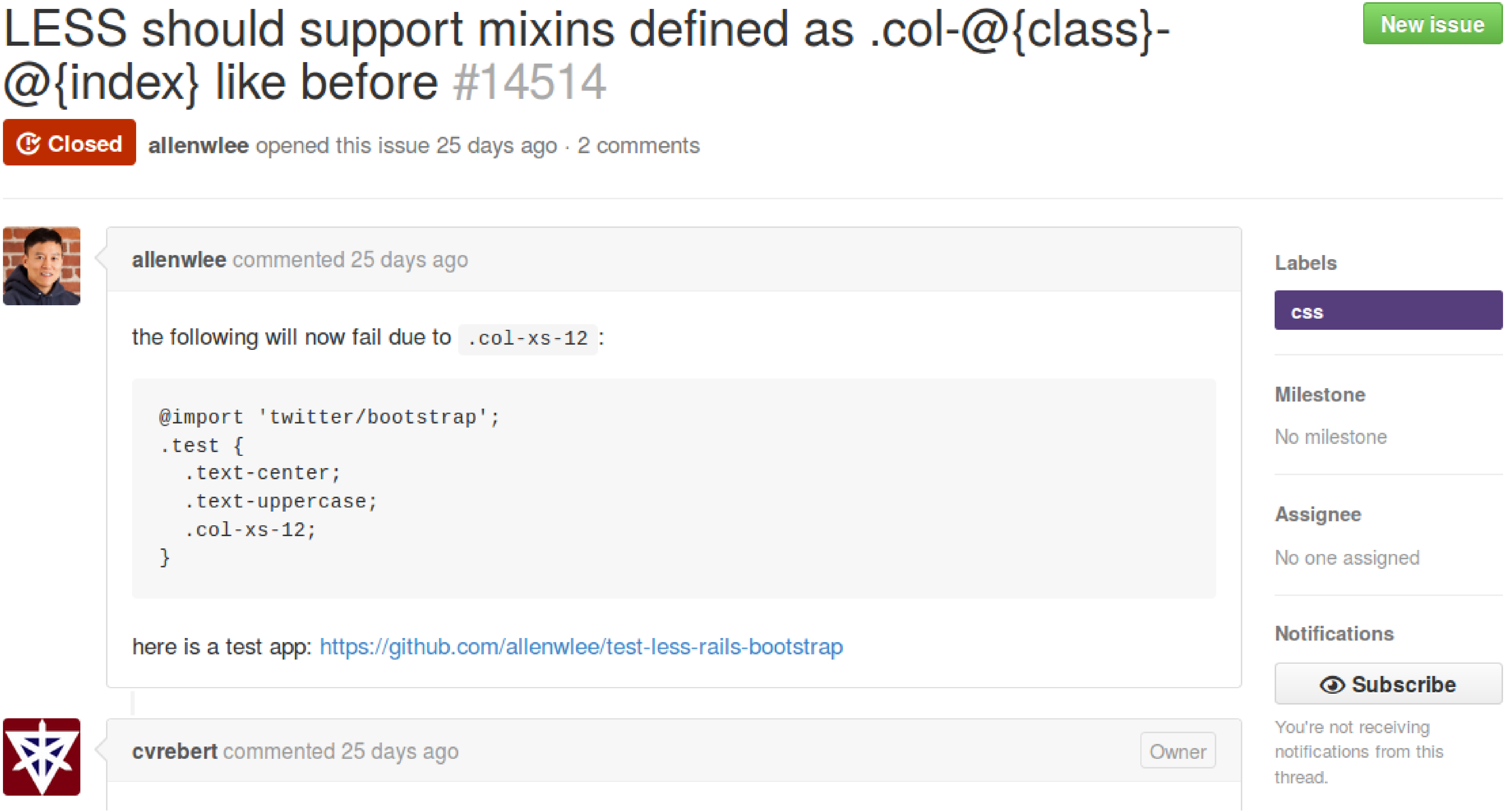
Missing border radius variables for small and large inputs css 💬 9
#14511 opened 26 days ago by hoho v3.2.1

has-feedback for RTL languages css feature 💬 2
#14510 opened 26 days ago by idleberg

Disabled fieldsets don't disable input fields on IE (11) css docs 💬 7

ub.com/twbs/bootstrap/issues?q=is:issue+is:closed

Data: Issue Trackers



LESS should support mixins defined as .col-@{class}-@{index} like before #14514

allenwlee opened this issue 25 days ago · 2 comments

allenwlee commented 25 days ago

the following will now fail due to `.col-xs-12`:

```
@import 'twitter/bootstrap';
.test {
  .text-center;
  .text-uppercase;
  .col-xs-12;
}
```

here is a test app: https://github.com/allenwlee/test-less-rails-bootstrap

cvrebert commented 25 days ago

Labels

css

Milestone

No milestone

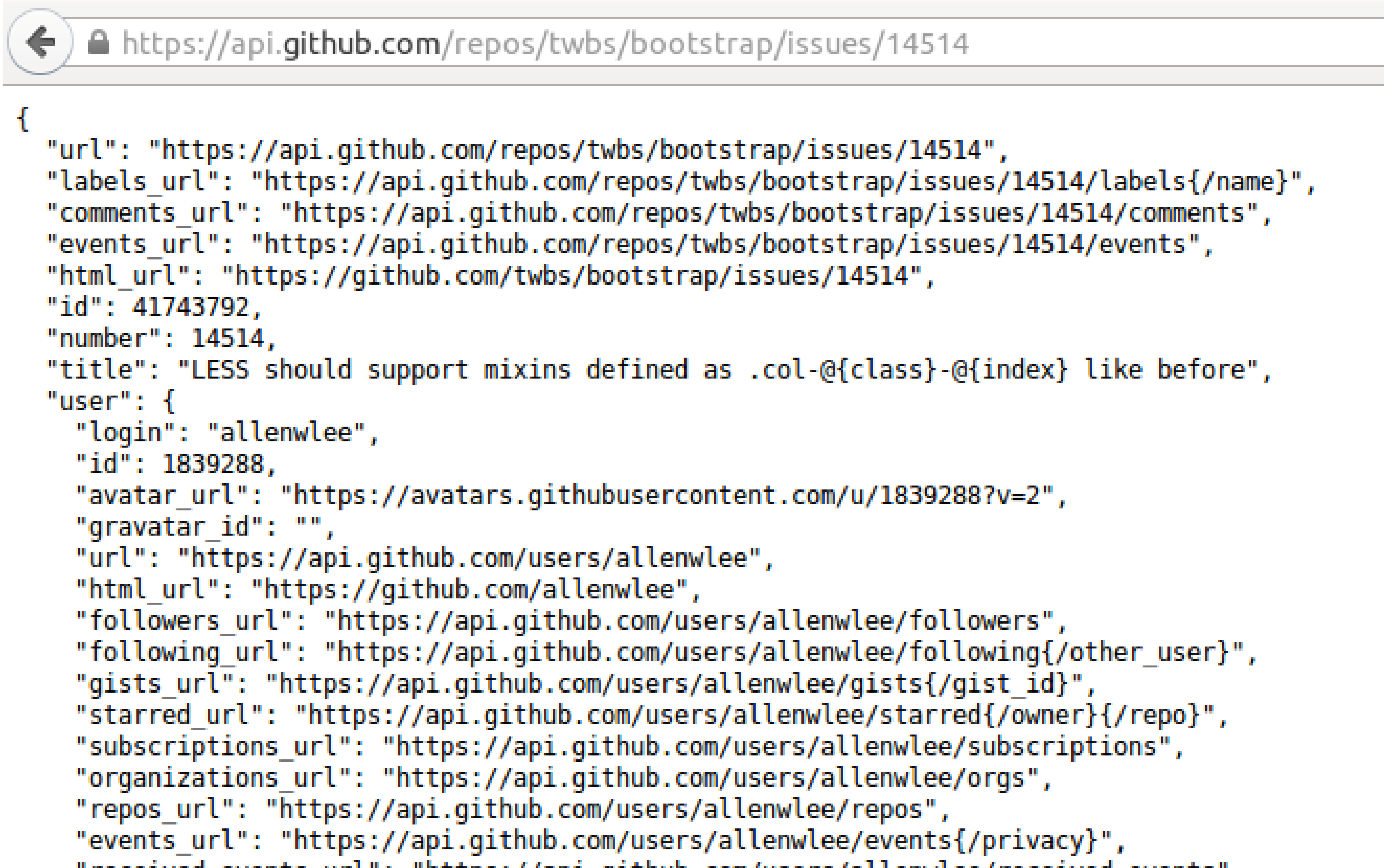
Assignee

No one assigned

Notifications

You're not receiving notifications from this thread.

Data: Issue Trackers



https://api.github.com/repos/twbs/bootstrap/issues/14514

```
{
  "url": "https://api.github.com/repos/twbs/bootstrap/issues/14514",
  "labels_url": "https://api.github.com/repos/twbs/bootstrap/issues/14514/labels{/name}",
  "comments_url": "https://api.github.com/repos/twbs/bootstrap/issues/14514/comments",
  "events_url": "https://api.github.com/repos/twbs/bootstrap/issues/14514/events",
  "html_url": "https://github.com/twbs/bootstrap/issues/14514",
  "id": 41743792,
  "number": 14514,
  "title": "LESS should support mixins defined as .col-@{class}-@{index} like before",
  "user": {
    "login": "allenwlee",
    "id": 1839288,
    "avatar_url": "https://avatars.githubusercontent.com/u/1839288?v=2",
    "gravatar_id": "",
    "url": "https://api.github.com/users/allenwlee",
    "html_url": "https://github.com/allenwlee",
    "followers_url": "https://api.github.com/users/allenwlee/followers",
    "following_url": "https://api.github.com/users/allenwlee/following{/other_user}",
    "gists_url": "https://api.github.com/users/allenwlee/gists{/gist_id}",
    "starred_url": "https://api.github.com/users/allenwlee/starred{/owner}{/repo}",
    "subscriptions_url": "https://api.github.com/users/allenwlee/subscriptions",
    "organizations_url": "https://api.github.com/users/allenwlee/orgs",
    "repos_url": "https://api.github.com/users/allenwlee/repos",
    "events_url": "https://api.github.com/users/allenwlee/events{/privacy}",
```

Octokit Issue Extractor

- Let's go look at github_issues_to_json.rb
- Uses github API
- Has to query multiple pages
- Needs config.json filled out with a real gh username and password
- https://bitbucket.org/abram/lda-chapter-tutorial

Go look at the code!

Issue Example

- Go and look at mirror-gh.sh
- Go and look at github_issues_to_json.rb
- Go and look at data/*/large.json

Pre-processing

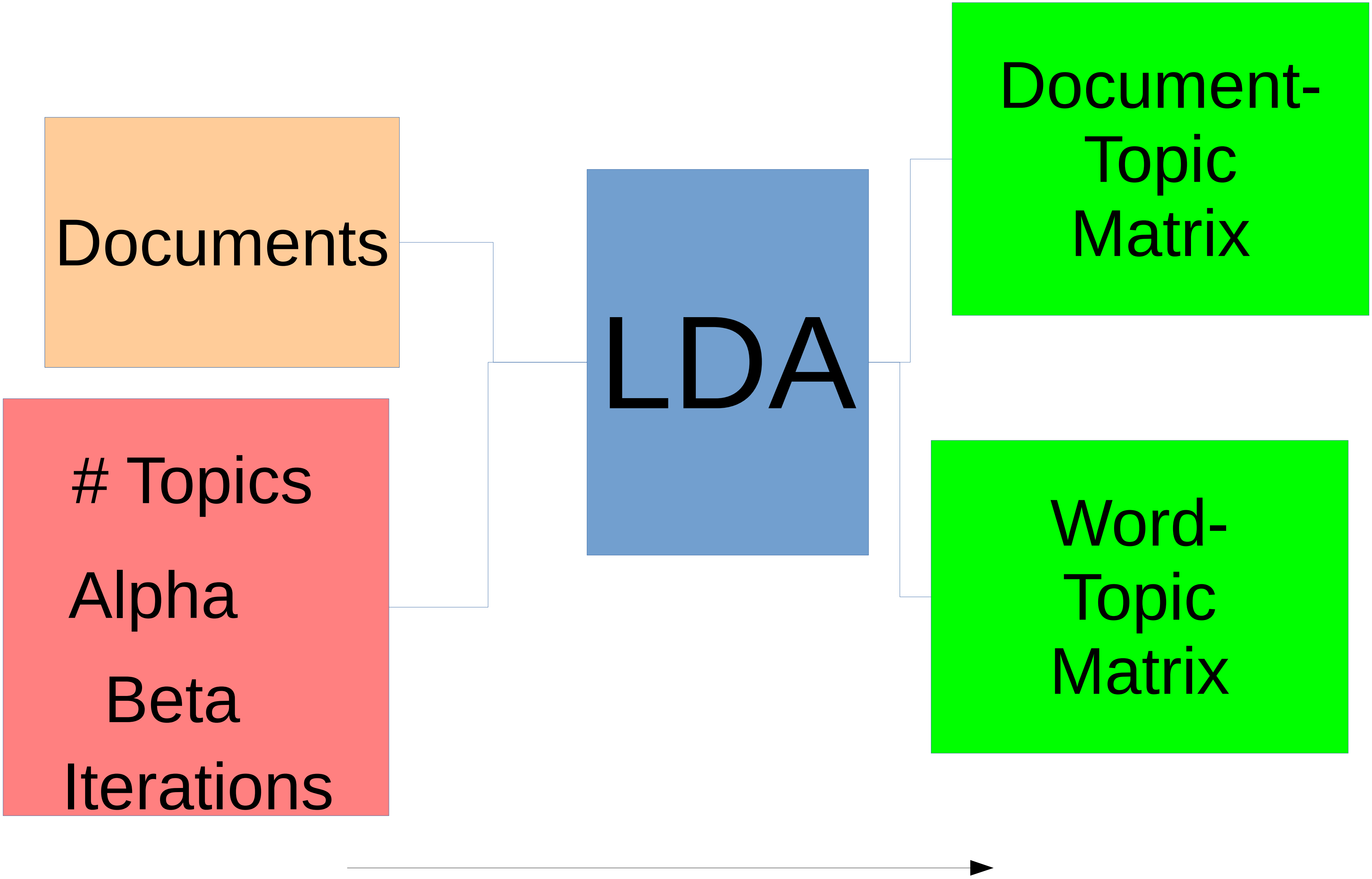
- Loading text
- Mapping text into final textual representation
- Lexical analysis of the text
- Optionally removing stop words
- Optionally stemming
- Building a vocabulary
- Optionally removing uncommon or very common words
- Mapping each text document into a word-bag

Example Preprocessing

From lda.py:

```
def tokenize( text, tokenizer=_tokenizer):
    tokens = filter_stopwords(
        tokenizer.tokenize( text.lower() ) )
    return tokens
```

LDA



Documents

# Topics

Alpha

Beta

Iterations

LDA

Document-Topic Matrix

Word-Topic Matrix

Alpha and Beta hyperparameters

- Actually vectors of parameters
- Most people use a constant setting
- A rule of thumb:
  - < 1/topic
- β is for topics: specific topics or not
- α is for documents: associated to few or many topics
- Larger values β lead to broad topics and smaller values of β lead to narrow topics
- If α is near 1, we expect to see documents with few topics and documents with many topics in equal proportion.
- If α is less than one, we expect most documents to only use a few topics.
- If α is greater than one, we expect most documents to use almost every topic.

In the demo: K Topics = 20, α = 0.01, β = 0.01

Parameter Tuning?

- Increasing topics increases memory use
  - But increasing the number of topics will often make you miss topics
- Joshua Campbell says use
  - Mallet or
  - Blei's C implementation

Run it!

- Run on existing data:

```
python lda_from_json.py --file \
data/boostrap/large.json --passes 10 \
 --alpha 0.01 --beta 0.01 --topics 20
```

- Or

```
bash project.sh bootstrap
```

Outputs!

- summary.json
  - JSON summary of the top topic words for each topic extracted, ranked by weight.
- document_topic_map.json
  - Document ID mapped to document topic matrix for that document
- document_topic_map.csv
  - unnormalized topic weights
- document_topic_map_norm.csv
  - Normalized topic weights

Spreadsheet example...

- Let's load the norm.csv file into libreoffice

Data

- Image: http://dub.softwareprocess.es/2014/LDA-Tutorial.ova
- Repo: https://bitbucket.org/abram/lda-chapter-tutorial/